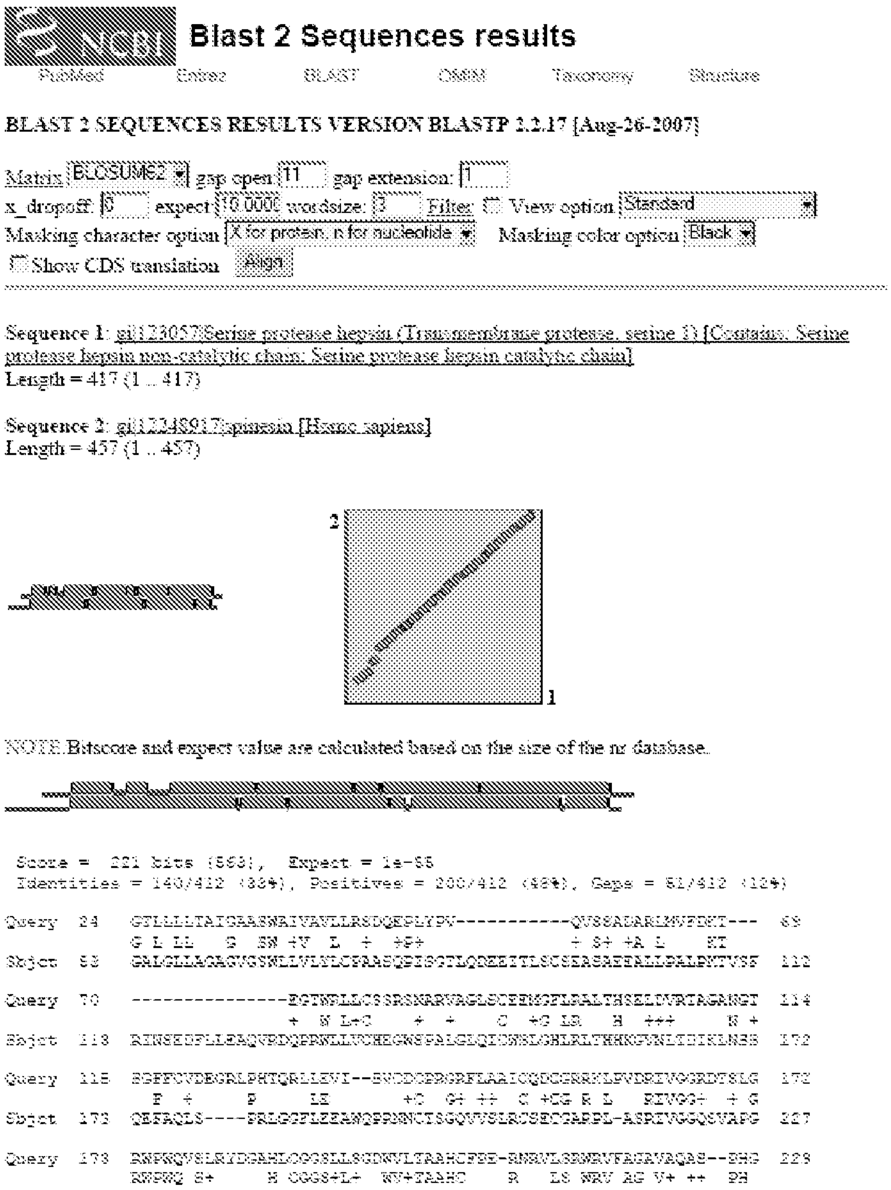
Appendix 1. BLAST alignments

A. hepsin vs spinesin:



Blast 2 Sequences results

BLAST 2 SEQUENCES RESULTS VERSION BLASTP 2.2.17 [Aug-26-2007]

Matrix BLOSUM62 ▾ gap open 11 gap extension 1
x_dropoff 50 expect 10.0000 wordsize 3 Filter □ View option Standard ▾
Masking character option X for protein, n for nucleotide ▾ Masking color option Black ▾
□ Show CDS translation Align

Sequence 1: gi|123057|Serine protease hepsin (Transmembrane protease, serine 1) [Contains: Serine protease hepsin non-catalytic chain; Serine protease hepsin catalytic chain]
Length = 417 (1 .. 417)

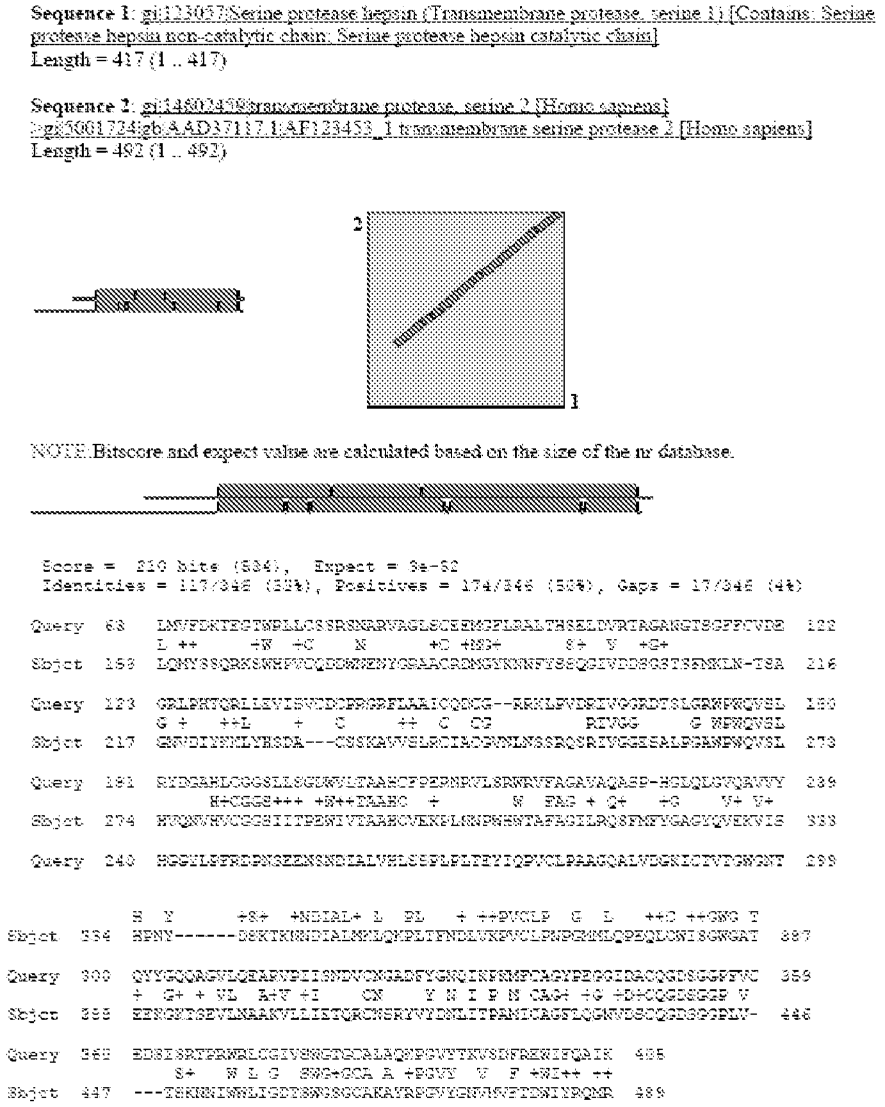
Sequence 2: gi|12248917|spinesin [Homo sapiens]
Length = 457 (1 .. 457)

NOTE:Bitscore and expect value are calculated based on the size of the nr database.

```
 Score =  221 bits (563),  Expect = 1e-55
 Identities = 140/412 (33%), Positives = 200/412 (48%), Gaps = 51/412 (12%)

Query  24   GTLLLLTAIGAASWAIVAVLLRSDQEPLYPV-----------QVSSADARLMVFDKT---  69
            G L LL    G  SW +V  L  +  +P+              + S+ +A L    KT
Sbjct  53   GALGLLAGAGVGSWLLVLYLCPAASQPISGTLQDEEITLSCSEASAEEALLPALPKTVSF  112

Query  70   ---------------EGTWRLLCSSRSNARVAGLSCEEMGFLRALTHSELDVRTAGANGT  114
                           +  W L+C    +  +    C  +G LR   H  +++     N +
Sbjct  113  RINSEDFLLEAQVRDQPRWLLVCHEGWSPALGLQICWSLGHLRLTHHKGVNLTDIKLNSS  172

Query  115  SGFFCVDEGRLPHTQRLLEVI--SVCDCPRGRFLAAICQDCGRRKLPVDRIVGGRDTSLG  172
              F  +     P     LE       +C  G+ ++  C +CG R L   RIVGG+  + G
Sbjct  173  QEFAQLS----PRLGGFLEEAWQPRNNCTSGQVVSLRCSECGARPL-ASRIVGGQSVAPG  227

Query  173  RWPWQVSLRYDGAHLCGGSLLSGDWVLTAAHCFPE-RNRVLSRWRVFAGAVAQAS--PHG  229
            RWPWQ S+     H CGGS+L+  WV+TAAHC    R   LS WRV AG V+ ++  PH
```

```
Sbjct  228  RWPWQASVALGFRHTCGGSVLAPRWVVTAAHCMHSFRLARLSSWRVHAGLVSHSAVRPHQ  287

Query  230  LQLGVQAVVYHGGYLPFRDPNSEENSNDIALVHLSSPLPLTEYIQPVCLPAAGQALVDGK  289
              L V+ ++ H  Y      +++ +  D+AL+ L + L  ++ +  VCLPA  Q    G
Sbjct  288  GAL-VERIIPHPLY------SAQNHDYDVALLRLQTALNFSDTVGAVCLPAKEQHFPKGS  340

Query  290  ICTVTGWGNTQ-YYGQQAGVLQEARVPIISNDVCNGADFYGNQIKPKMFCAGYPEGGIDA  348
             C V+GWG+T   +   + +LQ+  VP+ S  +CN +  Y   + P+M CAGY +G  DA
Sbjct  341  RCNVSGWGHTHPSHTYSSIMLQDTVVPLFSTQLCNSSCVYSGALTPRMLCAGYLDGRADA  400

Query  349  CQGDSGGPFVCEDSISRTPRWRLCGIVSWGTGCALAQKPGVYTKVSDFREWI  400
            CQGDSGGP VC D       WRL G+VSWG  CA    PGVY KV++F +WI
Sbjct  401  CQGDSGGPLVCPDG----DTWRLVGVVSWGRACAEPNHPGVYAKVAEFLDWI  448
```

B. hepsin vs TMPRSS2:



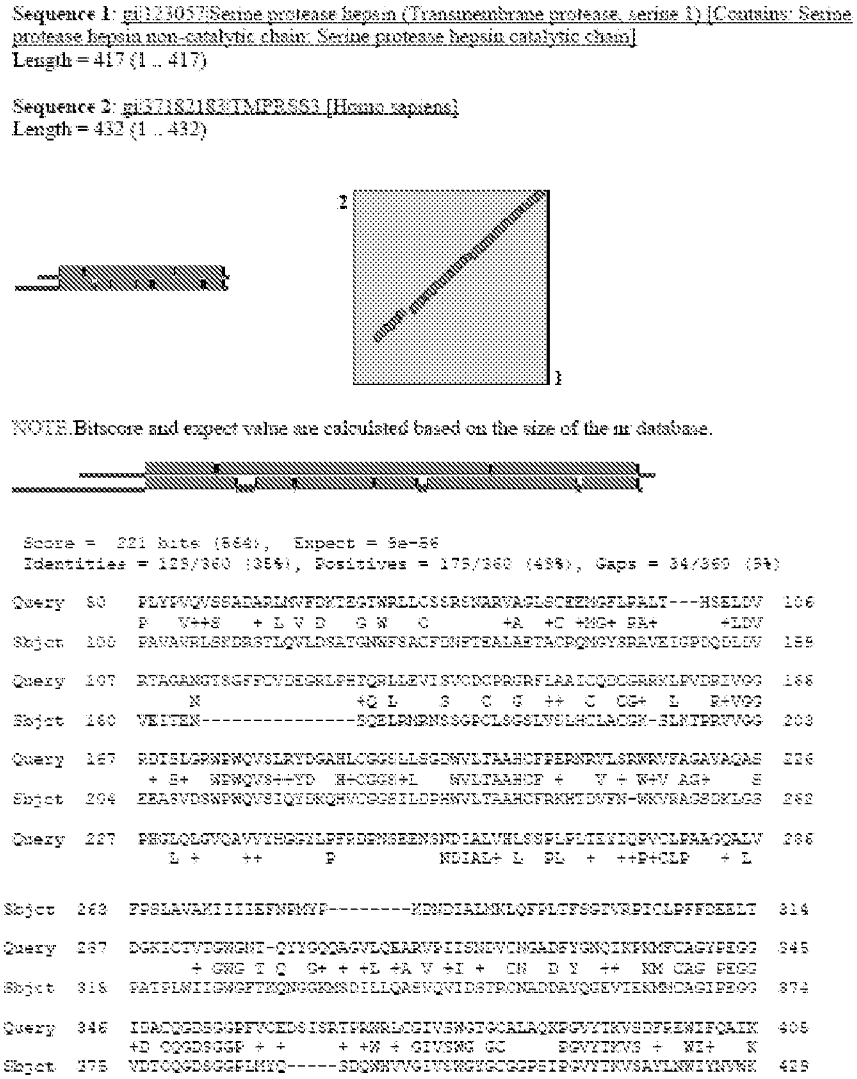
Sequence 1: gi|123057|Serine protease hepsin (Transmembrane protease, serine 1) [Contains: Serine protease hepsin non-catalytic chain; Serine protease hepsin catalytic chain]
Length = 417 (1 .. 417)

Sequence 2: gi|14602458|transmembrane protease, serine 2 [Homo sapiens]
>gi|5001724|gb|AAD37117.1|AF123453_1 transmembrane serine protease 2 [Homo sapiens]
Length = 492 (1 .. 492)

NOTE:Bitscore and expect value are calculated based on the size of the nr database.

```
 Score =  210 bits (534),  Expect = 3e-52
 Identities = 117/346 (33%), Positives = 174/346 (50%), Gaps = 17/346 (4%)

Query  63   LMVFDKTEGTWRLLCSSRSNARVAGLSCEEMGFLRALTHSELDVRTAGANGTSGFFCVDE  122
            L ++     +W  +C    N      +C +MG+      S+  V  +G+
Sbjct  153  LQMYSSQRKSWHPVCQDDWNENYGRAACRDMGYKNNFYSSQGIVDDSGSTSFMKLN-TSA  216

Query  123  GRLPHTQRLLEVISVCDCPRGRFLAAICQDCG--RRKLPVDRIVGGRDTSLGRWPWQVSL  180
            G +   ++L    +   C     ++  C  CG         RIVGG     G WPWQVSL
Sbjct  217  GNVDIYKKLYHSDA---CSSKAVVSLRCIACGVNLNSSRQSRIVGGESALPGAWPWQVSL  273

Query  181  RYDGAHLCGGSLLSGDWVLTAAHCFPERNRVLSRWRVFAGAVAQASP-HGLQLGVQAVVY  239
                 H+CGGS+++ +W++TAAHC  +       W  FAG + Q+   +G    V+ V+
Sbjct  274  HVQNVHVCGGSIITPEWIVTAAHCVEKPLNNPWHWTAFAGILRQSFMFYGAGYQVEKVIS  333

Query  240  HGGYLPFRDPNSEENSNDIALVHLSSPLPLTEYIQPVCLPAAGQALVDGKICTVTGWGNT  299


            H  Y      +S+  +NDIAL+ L  PL   + ++PVCLP  G  L   ++C ++GWG T
Sbjct  334  HPNY------DSKTKNNDIALMKLQKPLTFNDLVKPVCLPNPGMMLQPEQLCWISGWGAT  387

Query  300  QYYGQQAGVLQEARVPIISNDVCNGADFYGNQIKPKMFCAGYPEGGIDACQGDSGGPFVC  359
            +  G+ + VL  A+V +I    CN    Y N I P M CAG+ +G +D+CQGDSGGP V
Sbjct  388  EEKGKTSEVLNAAKVLLIETQRCNSRYVYDNLITPAMICAGFLQGNVDSCQGDSGGPLV-  446

Query  360  EDSISRTPRWRLCGIVSWGTGCALAQKPGVYTKVSDFREWIFQAIK  405
                S+   W L G  SWG+GCA A +PGVY  V  F +WI++ ++
Sbjct  447  ---TSKNNIWWLIGDTSWGSGCAKAYRPGVYGNVMVFTDWIYRQMR  489
```

C. hepsin vs TMPRSS3:



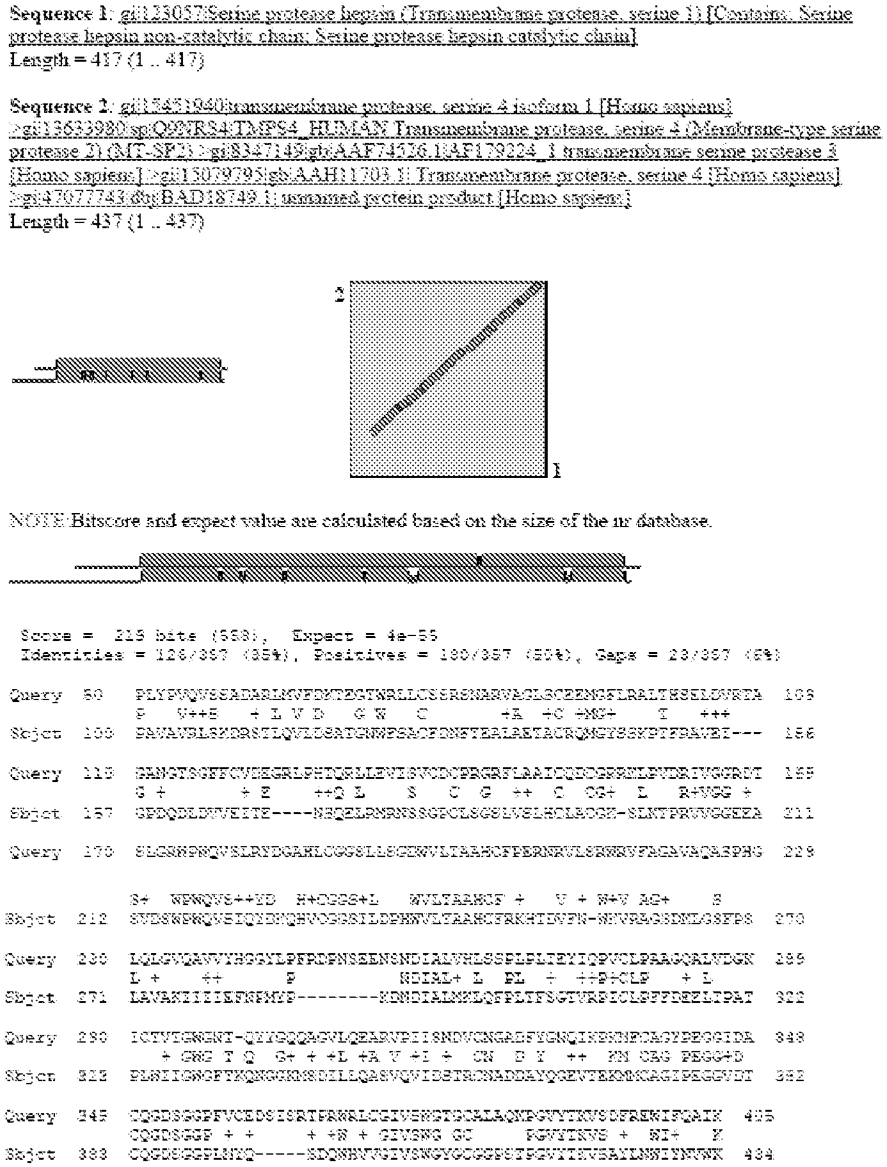
Sequence 1: gi|123057|Serine protease hepsin (Transmembrane protease, serine 1) [Contains: Serine protease hepsin non-catalytic chain; Serine protease hepsin catalytic chain]
Length = 417 (1 .. 417)

Sequence 2: gi|37182183|TMPRSS3 [Homo sapiens]
Length = 432 (1 .. 432)

NOTE:Bitscore and expect value are calculated based on the size of the nr database.

```
 Score =  221 bits (564),  Expect = 5e-56
 Identities = 125/360 (35%), Positives = 175/360 (48%), Gaps = 34/360 (9%)

Query  50   PLYPVQVSSADARLMVFDKTEGTWRLLCSSRSNARVAGLSCEEMGFLRALT---HSELDV  106
            P   V++S   + L V D   G W   C       +A  +C +MG+ RA+      +LDV
Sbjct  100  PAVAVRLSNDRSTLQVLDSATGNWFSACFDNFTEALAETACRQMGYSRAVEIGPDQDLDV  159

Query  107  RTAGANGTSGFFCVDEGRLPHTQRLLEVISVCDCPRGRFLAAICQDCGRRKLPVDRIVGG  166
                 N               +Q L    S   C  G  ++  C  CG+  L   R+VGG
Sbjct  160  VEITEN---------------SQELPNRNSSGPCLSGSLVSLHCLACGK-SLKTPRVVGG  203

Query  167  RDTSLGRWPWQVSLRYDGAHLCGGSLLSGDWVLTAAHCFPERNRVLSRWRVFAGAVAQAS  226
             + S+  WPWQVS++YD  H+CGGS+L   WVLTAAHCF +   V + W+V AG+    S
Sbjct  204  EEASVDSWPWQVSIQYDNQHVCGGSILDPHWVLTAAHCFRKHTDVFN-WKVRAGSDKLGS  262

Query  227  PHGLQLGVQAVVYHGGYLPFRDPNSEENSNDIALVHLSSPLPLTEYIQPVCLPAAGQALV  286
               L +    ++      P          NDIAL+ L  PL  +  ++P+CLP   + L

Sbjct  263  FPSLAVAKIIIIEFNFMYP--------KDNDIALMKLQFPLTFSGTVRPICLPFFDEELT  314

Query  287  DGKICTVTGWGNT-QYYGQQAGVLQEARVPIISNDVCNGADFYGNQIKPKMFCAGYPEGG  345
                  + GWG T Q  G+ + +L +A V +I +  CN  D Y  ++  KM CAG PEGG
Sbjct  315  PATPLWIIGWGFTKQNGGKMSDILLQASVQVIDSTRCNADDAYQGEVTEKMMCAGIPEGG  374

Query  346  IDACQGDSGGPFVCEDSISRTPRWRLCGIVSWGTGCALAQKPGVYTKVSDFREWIFQAIK  405
            +D CQGDSGGP + +     + +W + GIVSWG GC      PGVYTKVS +  WI+   K
Sbjct  375  VDTCQGDSGGPLMYQ-----SDQWHVVGIVSWGYGCGGPSTPGVYTKVSAYLNWIYNVWK  429
```

D. hepsin vs TMPRSS4:



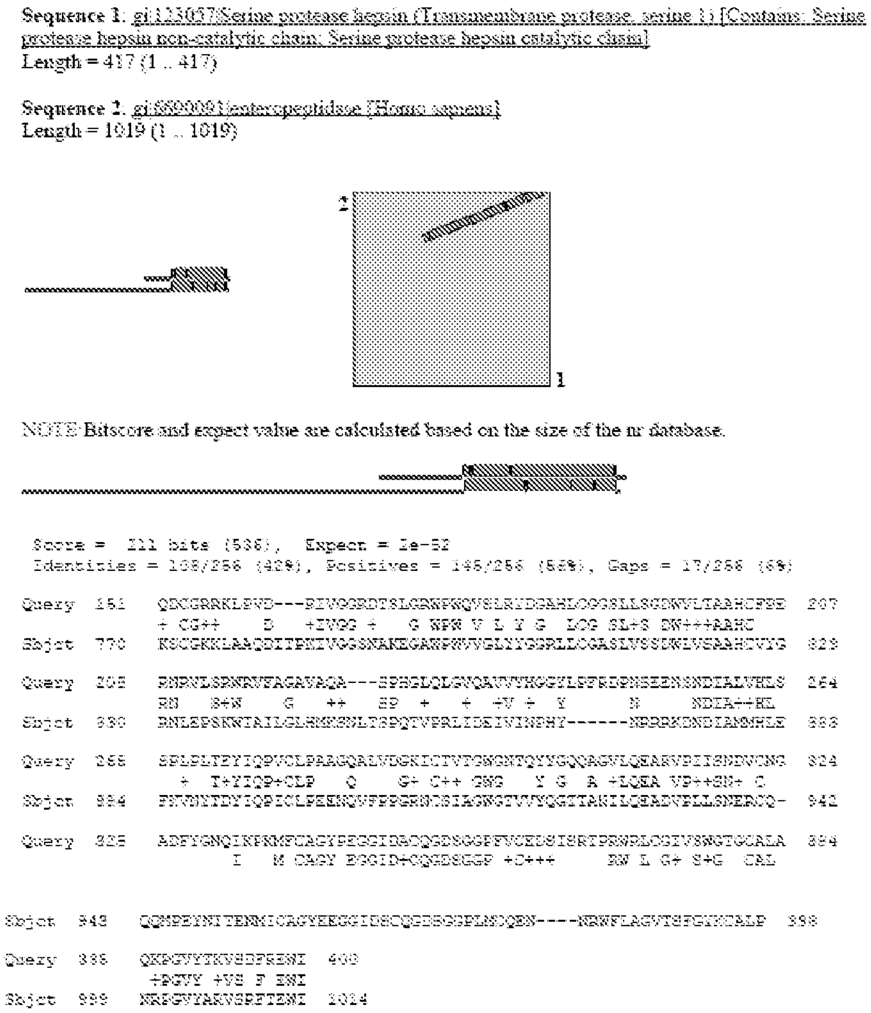
Sequence 1: gi|123057|Serine protease hepsin (Transmembrane protease, serine 1) [Contains: Serine protease hepsin non-catalytic chain; Serine protease hepsin catalytic chain]
Length = 417 (1 .. 417)

Sequence 2: gi|15451940|transmembrane protease, serine 4 isoform 1 [Homo sapiens] >gi|13633980|sp|Q9NRS4|TMPS4_HUMAN Transmembrane protease, serine 4 (Membrane-type serine protease 2) (MT-SP2) >gi|8347149|gb|AAF74526.1|AF179224_1 transmembrane serine protease 3 [Homo sapiens] >gi|15079795|gb|AAH11703.1| Transmembrane protease, serine 4 [Homo sapiens] >gi|47077743|dbj|BAD18749.1| unnamed protein product [Homo sapiens]
Length = 437 (1 .. 437)

NOTE:Bitscore and expect value are calculated based on the size of the nr database.

```
 Score =  215 bits (558),  Expect = 4e-55
 Identities = 126/357 (35%), Positives = 180/357 (50%), Gaps = 23/357 (6%)

Query  50   PLYPVQVSSADARLMVFDKTEGTWRLLCSSRSNARVAGLSCEEMGFLRALTHSELDVRTA  109
            P   V++S    + L V D   G W   C          +A  +C +MG+     T    +++
Sbjct  100  PAVAVRLSKDRSTLQVLDSATGNWFSACFDNFTEALAETACRQMGYSSKPTFRAVEI---  156

Query  110  GANGTSGFFCVDEGRLPHTQRLLEVISVCDCPRGRFLAAICQDCGRRKLPVDRIVGGRDT  169
            G +         + E    ++Q L    S   C  G  ++  C  CG+  L   R+VGG +
Sbjct  157  GPDQDLDVVEITE----NSQELRMRNSSGPCLSGSLVSLHCLACGK-SLKTPRVVGGEEA  211

Query  170  SLGRWPWQVSLRYDGAHLCGGSLLSGDWVLTAAHCFPERNRVLSRWRVFAGAVAQASPHG  229

           S+  WPWQVS++YD  H+CGGS+L   WVLTAAHCF +    V + W+V AG+     S
Sbjct  212  SVDSWPWQVSIQYDKQHVCGGSILDPHWVLTAAHCFRKHTDVFNWKVRAGSDKLGSFPS  270

Query  230  LQLGVQAVVYHGGYLPFRDPNSEENSNDIALVHLSSPLPLTEYIQPVCLPAAGQALVDGK  289
           L +    ++        P             NDIAL+ L  PL  +  ++P+CLP    + L
Sbjct  271  LAVAKIIIIEFNPMYP--------KDNDIALMKLQFPLTFSGTVRPICLPFFDEELTPAT  322

Query  290  ICTVTGWGNT-QYYGQQAGVLQEARVPIISNDVCNGADFYGNQIKPKMFCAGYPEGGIDA  348
               + GWG T Q  G+ + +L +A V +I +  CN  D Y  ++  KM CAG PEGG+D
Sbjct  323  PLWIIGWGFTKQNGGKMSDILLQASVQVIDSTRCNADDAYQGEVTEKMMCAGIPEGGVDT  382

Query  349  CQGDSGGPFVCEDSISRTPRWRLCGIVSWGTGCALAQKPGVYTKVSDFREWIFQAIK  405
           CQGDSGGP + +       + +W + GIVSWG GC        PGVYTKVS +   WI+    K
Sbjct  383  CQGDSGGPLMYQ-----SDQWHVVGIVSWGYGCGGPSTPGVYTKVSAYLNWIYNVWK  434
```

E. hepsin vs enteropeptidase:



Sequence 1: gi|123057|Serine protease hepsin (Transmembrane protease, serine 1) [Contains: Serine protease hepsin non-catalytic chain; Serine protease hepsin catalytic chain]
Length = 417 (1 .. 417)

Sequence 2: gi|5690091|enteropeptidase [Homo sapiens]
Length = 1019 (1 .. 1019)

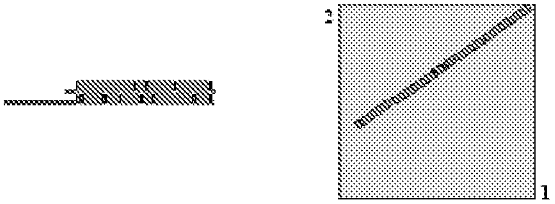
NOTE:Bitscore and expect value are calculated based on the size of the nr database.

```
 Score =  211 bits (536),  Expect = 1e-52
 Identities = 108/256 (42%), Positives = 145/256 (56%), Gaps = 17/256 (6%)

Query  151   QDCGRRKLPVD---RIVGGRDTSLGRWPWQVSLRYDGAHLCGGSLLSGDWVLTAAHCFPE  207
             + CG++    D   +IVGG +   G WPW V L Y G  LCG SL+S DW++++AAHC
Sbjct  770   KSCGKKLAAQDITPKIVGGSNAKEGAWPWVVGLYYGGRLLCGASLVSSDWLVSAAHCVYG  829

Query  208   RNRVLSRWRVFAGAVAQA---SPHGLQLGVQAVVYHGGYLPFRDPNSEENSNDIALVHLS  264
             RN   S+W    G   ++   SP  +   +  +V + Y       N     NDIA++HL
Sbjct  830   RNLEPSKWTAILGLHMKSNLTSPQTVPRLIDEIVINPHY------NRRRKDNDIAMMHLE  883

Query  265   SPLPLTEYIQPVCLPAAGQALVDGKICTVTGWGNTQYYGQQAGVLQEARVPIISNDVCNG  324
               +  T+YIQP+CLP   Q     G+ C++ GWG   Y G  A +LQEA VP++SN+ C
Sbjct  884   FKVNYTDYIQPICLPEENQVFPPGRNCSIAGWGTVVYQGTTANILQEADVPLLSNERCQ-  942

Query  325   ADFYGNQIKPKMFCAGYPEGGIDACQGDSGGPFVCEDSISRTPRWRLCGIVSWGTGCALA  384
                   I   M CAGY EGGID+CQGDSGGP +C+++      RW L G+ S+G  CAL

Sbjct  943   QQMPEYNITENMICAGYEEGGIDSCQGDSGGPLMCQEN----NRWFLAGVTSFGYKCALP  998

Query  385   QKPGVYTKVSDFREWI  400
              +PGVY +VS F EWI
Sbjct  999   NRPGVYARVSRFTEWI  1014
```

F. hepsin vs MSPL:

Sequence 1: gi|123057|Serine protease hepsin (Transmembrane protease, serine 1) [Contains: Serine protease hepsin non-catalytic chain; Serine protease hepsin catalytic chain]
Length = 417 (1 .. 417)

Sequence 2: gi|116256363|transmembrane protease, serine 13 [Homo sapiens]
Length = 567 (1 .. 567)



NOTE:Bitscore and expect value are calculated based on the size of the nr database.

```
 Score =  231 bits (588),  Expect = 1e-58
 Identities = 134/372 (36%), Positives = 196/372 (52%), Gaps = 29/372 (7%)

Query  39   IVAVLLRSDQEPLYPVQVSSADARLMVFDKTEGTWRLLCSSRSNARVAGLSCEEMGFLRA  98
            +V   L+SD+  L  V+     + L ++  +   W  +CSS  N   +  +C+++GF  A
Sbjct  214  VVDCKLKSDE--LGCVRFDWDKSLLKIYSGSSHQWLPICSSNWNDSYSEKTCQQLGFESA  271

Query  99   LTHSELDVRTAGANGTSGFFCVDEGRLPHTQRLLEVISVCDCPRGRFLAAICQDCGRRKL  158
               +E+  R          F      L +   +  E +   +CP  R+++  C  CG R +
Sbjct  272  HRTTEVAHRD---------FANSFSILRYNSTIQESLHRSECPSQRYISLQCSHCGLRAM  322

Query  159  PVDRIVGGRDTSLGRWPWQVSLRYDGAHLCGGSLLSGDWVLTAAHC-FPERNRVLSRWRV  217
               RIVGG   S  +WPWQVSL +   H+CGG+L+   WVLTAAHC F  R +VL  W+V
Sbjct  323  -TGRIVGGALASDSKWPWQVSLHFGTTHICGGTLIDAQWVLTAAHCFFVTREKVLEGWKV  381

Query  218  FAGAVAQASPHGL--QLGVQAVVYHGGYLPFRDPNSEENSNDIALVHLSSPLPLTEYIQP  275
            +AG    ++ H L     +  ++ +  Y        EE+  DIAL+ LS PL L+ +I P


Sbjct  382  YAGT---SNLHQLPEAASIAEIIINSNY------TDEEDDYDIALMRLSKPLTLSAHIHP  432

Query  276  VCLPAAGQALVDGKICTVTGWGNTQYYGQQAG-VLQEARVPIISNDVCNGADFYGNQIKP  334
             CLP  GQ     + C +TG+G T+    +    L+E +V +I    CN    Y + + P
Sbjct  433  ACLPMHGQTFSLNETCWITGFGKTRETDDKTSPFLREVQVNLIDFKKCNDYLVYDSYLTP  492

Query  335  KMFCAGYPEGGIDACQGDSGGPFVCEDSISRTPRWRLCGIVSWGTGCALAQKPGVYTKVS  394
            +M CAG   GG D+CQGDSGGP VCE    +  RW L G+ SWGTGC    KPGVYTKV+
Sbjct  493  RMMCAGDLRGGRDSCQGDSGGPLVCE----QNNRWYLAGVTSWGTGCGQRNKPGVYTKVT  548

Query  395  DFREWIFQAIKT  406
            +   WI+  +++
Sbjct  549  EVLPWIYSKMES  560
```